Cisco Nexus 9500 Series Switches Architecture

White Paper

November 2013

Contents

Nexus 9500 Series Switches Introduction 3

Scalable Control Plane on the Cisco Nexus 9500 Series Switches 5

- Supervisor Engine 5
- System Controllers 6

Non-Blocking Distributed Data Plane on the Cisco Nexus 9500 Series Switches 7

- Nexus 9500 Series Fabric Module 8
- Nexus 9500 Series Switches Line Card Architecture 9
- 36x 40GE QSFP Line Card (N9K-X9636PQ) 9
- 48x 1/10G SFP+ Line Card (N9K-X9564PX) 10
- 48x 1/10G BastT Line Card (N9K-X9564TX) 10

Nexus 9500 Series Unicast Packet Forwarding 11

- 1. Ingress Processing Pipeline 12
- 2. Fabric Module LPM Lookup 13
- 3. Egress Processing Pipeline 13

Nexus 9500 Series Multicast Packet Forwarding 14

Cisco QSFP Bi-Di Technology for 40Gbps Migration 15

Conclusion 16

Appendix 17

Nexus 9500 Series Switches Introduction

The Cisco Nexus 9500 Series is a family of modular switches that delivers industry leading high-performance, high-density and low-latency 1, 10, 40, and, in the future, 100 Gigabit Ethernet connectivity. Nexus 9500 Series Switches can operate in both the Application Centric Infrastructure (ACI) mode and the classic NX-OS mode. When they operate in the ACI mode, Nexus 9500 Series Switches form the foundation of the transformational ACI architecture for the fully integrated and automated network fabric solution driven by Application Network Profile. While running in the classic NX-OS mode, Nexus 9500 Series Switches are the first of its kind switches for highly scalable and high performance data center access and aggregation layers, with enhanced automation and programmability functionalities. This white paper focuses on the common hardware architecture of Nexus 9500 Series Switches and the packet forwarding implementation in the classic NX-OS mode.

The 8-slot Nexus 9508 switch (Figure 1) is the first available platform in the family that will be followed by the 4-slot and 16-slot platforms. The Cisco Nexus 9508 switch supports up to 1152 10GE ports or 288 40GE ports. The Cisco Nexus 9516 switch will double the port densities. Nexus 9500 Series Switches also provide high port density for 1G SFP/1GBase-T and 10G SFP+/10GBaseT connectivity. With the various chassis form factors, different line card types and flexible front panel port speeds, the Cisco Nexus 9500 Series provide superior networking solutions for small, medium and large mission critical data centers.

Figure 1. Cisco Nexus 9508 Switch

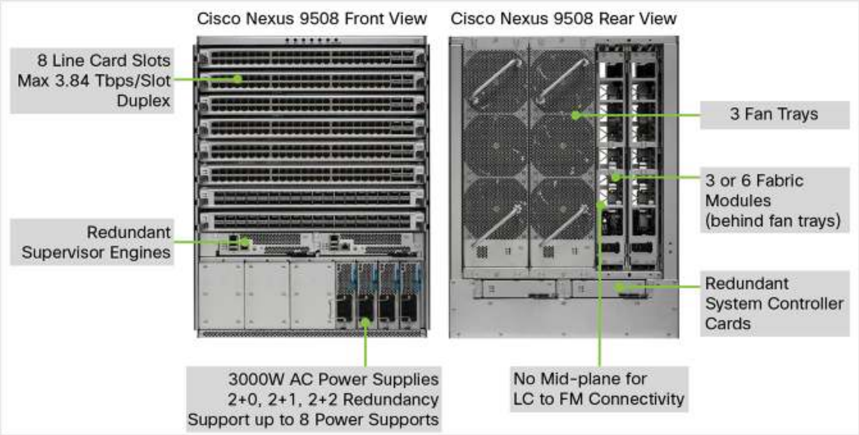


Table 1. Cisco Nexus 9500 Chassis and Forwarding Characteristics

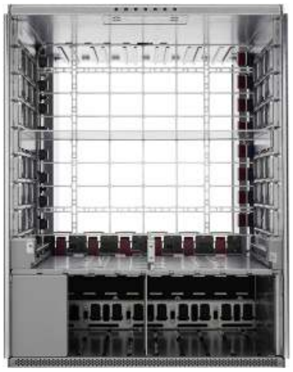
Metric	NEXUS 9504	NEXUS 9508	NEXUS 9516
Height	7 RU	13 RU	20 RU
Supervisor Slots	2	2	2
Fabric Module Slots	6	6	6
Line Card Slots	4	8	16
Max Fabric BW per Slot (Tbps)	3.84 Tbps	3.84 Tbps	3.84 Tbps
Max Fabric BW per System (Tbps)	15 Tbps	30 Tbps	60 Tbps
Max 1/10/40/ports	192/576/144	384/1152/288	768/2304/576
Max Forwarding Throughput per Line Card (Tbps)	2.88 Tbps	2.88 Tbps	2.88Tbps
Max Forwarding Throughput per System (Tbps)	11.52 Tbps	23.04 Tbps	46.08 Tbps
Air Flow	Front-to-Back	Front-to-Back	Front-to-Back
Power Supplies	4 x 3KW AC PSUs	8 x 3KW AC PSUs	8 x 3KW AC PSUs

Metric	NEXUS 9504	NEXUS 9508	NEXUS 9516
Fan Trays	3	3	3

The Cisco Nexus 9500 Series switches have a modular architecture that consists of switch chassis, supervisors, system controllers, fabric modules, line cards, power supplies and fan trays. Among these parts, supervisors, system controllers, line cards and power supplies are common components that can be shared among the entire Nexus 9500 product family.

The chassis of the Cisco Nexus 9500 Series has an innovative midplane-free design (Figure 2). A midplane is commonly used in modular platforms as a way to provide connectivity between line cards and fabric modules. Being an extra piece of hardware inside the switch chassis, it obstructs the cooling airflow. Hence, additional methods need to be applied to facilitate an airflow path, e.g. cut-outs on the midplane or airflow redirection, which result in a reduced cooling efficiency. Nexus 9500 Series is the industry-first switch platform that eliminates the need for a chassis midplane. With a precise alignment mechanism, Nexus 9500 Series switch line cards and fabric modules directly attach to each other with connecting pins. Line cards and fabric modules have the orthogonal orientations in the chassis so that each fabric module is connected to all line cards and vice versa. Without a midplane blocking the airflow path, the chassis design delivers a maximized cooling efficiency. It also allows a compact chassis design without the need for large cooling fans.

Figure 2. Nexus 9500 Midplane-free Chassis Design



Midplane-free chassis design greatly simplifies the switch platform deployment and hardware upgrade. In some cases where new components, such as new line cards or new fabric modules, are introduced, an upgrade to the midplane is required. This introduces complexity and more service disruption to the hardware upgrade process. The Cisco Nexus 9500 Series alleviate the need of midplane installation or upgrade. Another advantage of removing the midplane is significantly improved mean-time-to-repair. With a midplane, if you bend a pin on the midplane, the entire switch must be taken out of commission and disassembled to replace that midplane. With the 9500, the components that are damaged can be replaced without taking the other components of the chassis out of service.

In addition to the highest cooling efficiency, the Cisco Nexus 9500 Series also leads with high power efficiency. Its power supplies carry the industry standard 80PLUS Platinum certification for high efficiency. Line cards and fabric modules of Nexus 9500 Series are designed with a minimal number of ASICs that reduces the amount of heat pockets on a module. The results of these innovations is an unmatched, lowest power consumption per port:

Power Consumption/Port	10Gbps Port	40Gbps Port
Watt per Port	3.85W/port	15.4W/port

Scalable Control Plane on the Cisco Nexus 9500 Series Switches

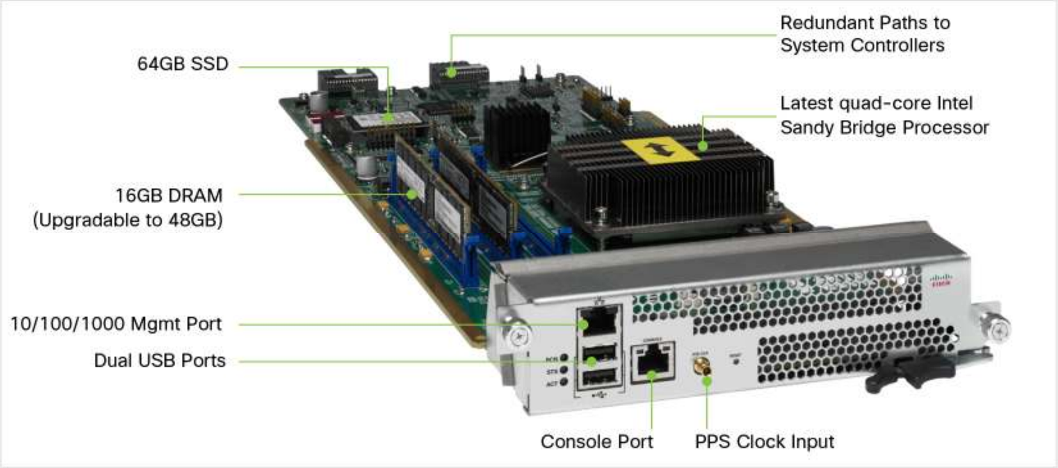
The Cisco Nexus 9500 supervisor engine delivers a scalable control plane for the Cisco Nexus 9500 Series Switches. The system controller offloads the internal component connectivity and management functions from the supervisor engine. Decoupling the internal management tasks from the supervisor engine increases the reliability of the switch control plane. It provides better modularity and resiliency of the entire switch system.

Supervisor Engine

The Cisco Nexus 9500 Series support redundant half-width supervisor engines that are responsible for control-plane functions. The switch software, Enhanced NX-OS, runs on the supervisor modules. The redundant supervisor modules take active and standby roles supporting stateful switch over in the event of supervisor module hardware failure, and In-Service Software Upgrade (ISSU), allowing software upgrade/maintenance without impacting production services.

The CPU complex of the Nexus 9500 supervisor is based on the Intel Romley platform with 4 core Sandy Bridge Exon processors. The default system memory size is 16 GB that is field upgradable to 48 GB. There is a built-in 64 GB SSD to provide additional on-board non-volatile storage. The high speed multi-core CPU and the large memory build the foundation for a fast and reliable control plane for the switch system. Control plane protocols will benefit from the ample computation horsepower and achieve fast initiation and instantaneous convergence upon network state changes. Additionally, the expandable large DRAM and multi-core CPU provides sufficient compute power and resources to support c-group based Linux containers in which third party applications can be installed and run in a well-contained environment. The on-board SSD provides extra storage for logs, image files and third party applications.

Figure 3. Cisco Nexus 9500 Supervisor Engine


Redundant Paths to System Controllers
64GB SSD
Latest quad-core Intel Sandy Bridge Processor
16GB DRAM (Upgradable to 48GB)
10/100/1000 Mgmt Port
Dual USB Ports
Console Port
PPS Clock Input

Supervisor Module	
Processor	Romley, 1.8GHz, 4 core
System Memory	16GB, upgradable to 48GB
RS-232 Serial Ports	One (RJ-45)
10/100/1000 Management Ports	One (RJ-45)
USB 2.0 Interface	Two
SSD Storage	64GB

The supervisor engine has a serial console port (RJ-45) and a 10/100/1000 Ethernet management port (RJ-45) for out-of-band management. Two USB 2.0 interfaces are supported using an external USB flash storage for image, syslog, configuration file transfer and other uses. A Pulse-per-second (PPS) clock input port on the supervisor module supports accurate timing synchronization.

The communications between the supervisor and the fabric modules or line cards utilize either Ethernet Out-of-Band Channel (EOBC) or Ethernet Protocol Channel (EPC). Both channels have a central hub on the System Controllers providing redundant paths to the System Controllers.

System Controllers

The System Controllers of Cisco Nexus 9500 Series are used to offload the internal non-data-path switching and management functions from the supervisor engines. It also provides the pathway for access to the power supplies and fan trays.

The System Controllers are the intra-system communication central switches. It hosts two main control and management communication paths, Ethernet Out-of-Band Channel (EOBC) and Ethernet Protocol Channel (EPC), between supervisor engines, line cards and fabric modules.

All intra-system management communication across modules takes place through the EOBC channel. The EOBC channel is provided via a switch chipset on the System Controllers that inter-connects all modules together, including supervisor engines, fabric modules and line cards.

The EPC channel handles intra-system data plane protocol communication. This communication pathway is provided by another redundant Ethernet switch chipset on the System Controllers. Unlike the EOBC channel, the EPC switch only connects fabric modules to supervisor engines. If protocol packets need to be sent to the supervisors, line cards utilize the internal data path to transfer packets to fabric modules. The fabric modules then redirect the packet via the EPC channel to the supervisor engines.

The System Controller also communicates with and manages power supply units and fan controllers via the redundant system management bus (SMB).

The Cisco Nexus 9500 Series supports redundant System Controllers. When two System Controllers are present in a chassis, an arbitration process will select the active System Controller. The other one will assume the secondary or standby role to provide redundancy.

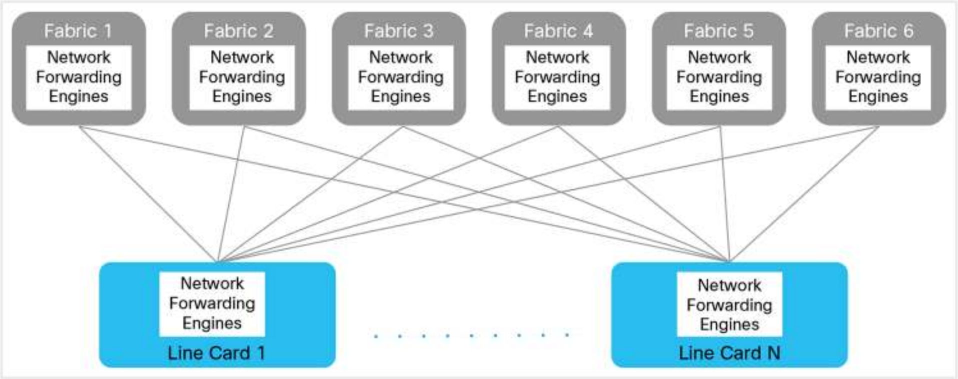
Figure 4. Cisco Nexus 9500 Series System Controller

Non-Blocking Distributed Data Plane on the Cisco Nexus 9500 Series Switches

While the switch control plane is centrally run on the supervisor engines, the packet lookup and forwarding functions in the data plane are conducted in a highly distributed fashion involving line cards and fabric modules.

Both line cards and fabric modules of the Cisco Nexus 9500 Series are equipped with multiple network forwarding engines (NFE) that perform packet lookup, processing and forwarding functions. The Nexus 9500 Series Switches are designed bearing in mind a non-blocking architecture and full line-rate performance on all ports, independent of packet size. Since many modern-day datacenter applications use small-sized packets, it is essential to support line-rate performance for even the smallest packet of 64-bytes. In order to achieve this level of forwarding capability, Nexus 9500 Series line cards and fabric modules are architected with the required number of NFEs. Up to 24 40GE ports are used on each NFE to guarantee line-rate performance. Among the 24 40GE ports, 12 40GE ports, clocked at 42GE to accommodate the extra bits in the internal frame header, are used for internal connectivity towards fabric modules. The other 12 ports are used as front panel interfaces to support 1, 10, 40 and future 100GE user data ports.

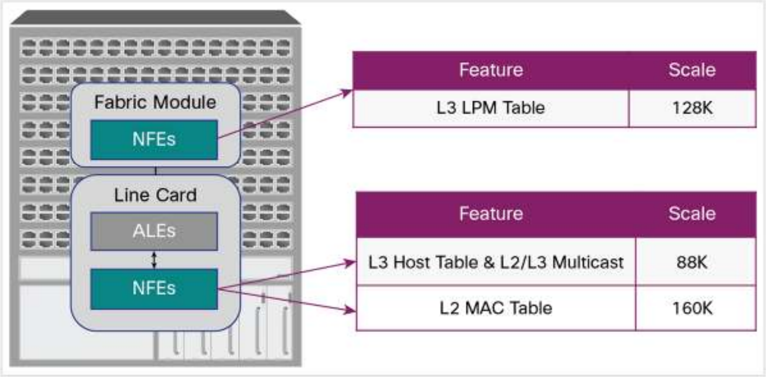
Figure 5. Distributed Data Plane of Nexus 9500 Series Switches


Fabric 1
Network Forwarding Engines
Fabric 2
Network Forwarding Engines
Fabric 3
Network Forwarding Engines
Fabric 4
Network Forwarding Engines
Fabric 5
Network Forwarding Engines
Fabric 6
Network Forwarding Engines
Network Forwarding Engines
Line Card 1
Network Forwarding Engines
Line Card N

The network forwarding engines use a combination of dedicated TCAM table space and shared hash table memory known as Unified Forwarding Table (UFT) to store Layer-2 and Layer-3 forwarding information. The UFT can be flexibly partitioned into three forwarding tables, the MAC Address Table, the IP Host Table and the LPM Table. This programmable memory sharing approach provides flexibility to cater to different deployment scenarios and increases the efficiency of the memory resource utilization.

To maximize the system-wide forwarding scalability, the Nexus 9500 Series switches are designed to use the UFT tables on line cards and fabric modules for different forwarding lookup functions. The UFT on line cards stores L2 MAC table and L3 Host table. Therefore line cards are responsible for L2 switching lookup and L3 host routing lookup. The UFT on fabric modules hosts L3 LPM table, and perform L3 LPM routing lookup. Both line cards and fabric modules have multicast tables and take part in distributed multicast lookup and packet replication. Multicast shares the same table resource with L3 Host entries on line cards. Figure 6 depicts the system-wide forwarding scalability of Nexus 9500 Series switches.

Figure 6. Nexus 9500 System-wide Forwarding Scalability



Feature	Scale
L3 LPM Table	128K

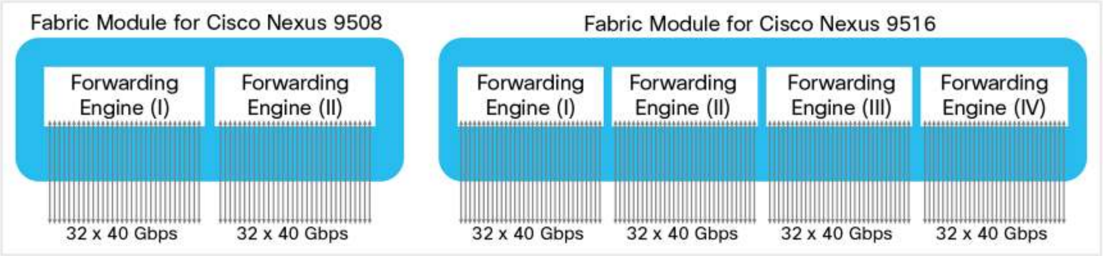
Feature	Scale
L3 Host Table & L2/L3 Multicast	88K
L2 MAC Table	160K

Nexus 9500 Series Fabric Module

A Nexus 9500 Series switch can have up to six fabric modules that all function in active mode. Each fabric module consists of multiple network forwarding engines (NFE), 2 for a Nexus 9508 Switch and 4 for a Nexus 9516 Switch (Figure 7).

For a Nexus 9508 switch, up to twelve NFEs can be available on its fabric models. This provides the required data path bandwidth and packet forwarding capacity to achieve a true non-blocking architecture. Hence the Nexus 9508 can support real line-rate performance, independent of packet size, on all line cards.

Figure 7. Nexus 9500 Series Fabric Module



The Fabric Module of Nexus 9500 Series Switches performs the following important functions in the modular chassis architecture:

- Provide high-speed non-blocking data forwarding connectivity for line cards. All links on network forwarding engines are active data paths. Each fabric module can provide up to 8 40Gbps links to every line card slot. A nexus 9500 chassis deployed with 6 fabric modules can potentially provide 48 40Gbps fabric paths to each line card slot. This is equivalent to 3.84 Tbps full duplex bandwidth per slot.
- Perform distributed LPM (Longest Prefix Match) routing lookup for IPv4 and IPv6 traffic. LPM forwarding information is stored on fabric modules on a Nexus 9500 Series switch. It supports up to 128K IPv4 prefixes or 32K IPv6 prefixes.
- Perform distributed multicasts lookup and packet replication to send copies of multicast packets to receiving egress NFEs.

Nexus 9500 Series Switches Line Card Architecture

A Nexus 9500 Series Switches line card can be categorized into two types, aggregation line cards and Application Centric Infrastructure (ACI) ready leaf line lards. Aggregation line cards provide high-density 10GE/40GE connectivity on a Nexus 9500 switch that runs in the classic NX-OS mode. ACI Ready Leaf Line Cards can operate in both the classis NX-OS mode and ACI mode.

All Nexus 9500 line cards consist of multiple NFEs for packets lookup and forwarding. Additionally, the ACI ready leaf line cards have a set of Application Leaf Engines (ALE). At its name indicates, ALE performs ACI leaf node functions when the Nexus 9500 switch is deployed as a leaf node in an ACI infrastructure. When the Nexus 9500 switch is operating in the classic NX-OS mode, the ALE on an ACI ready leaf line card mainly provides additional buffering and facilitates some networking functions, such as routing within a VxLAN overlay.

NFEs on a line card do L2 switching lookups and L3 host routing lookups. Line cards are equipped with various numbers of NFEs to support full line-rate forwarding performance for all IP packet sizes on all front panel ports.

In addition to the line-rate data plane performance, Nexus 9500 Series switch line cards also have a built-in dual-core CPU. This CPU is used to offload or speed up some control plane tasks such as programming the hardware table resources, collecting and sending line card counters and statistics, and offloading BFD protocol handling from the supervisors. This provides significant improvement to the system control plane performance.

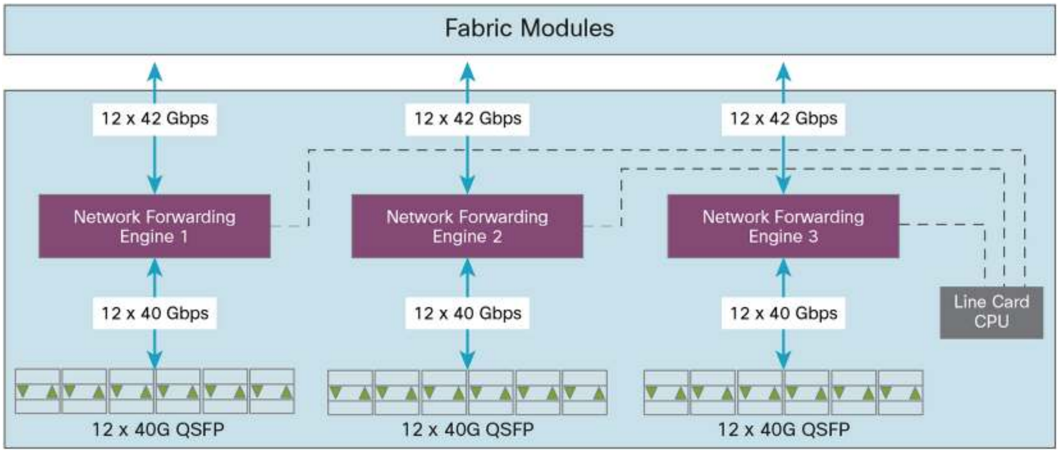
36x 40GE QSFP Line Card (N9K-X9636PQ)

N9K-X9636PQ (Figure 8) is an aggregation line card providing 36 40GE QSFP front panel ports. It has three network forwarding engines for packet forwarding, each supporting 12 40GE front panel ports and 12 internal ports to the fabric modules (clocked at 42 Gbps rate to accommodate the internal frame overhead). All of the 36 40GE front panel ports on N9K-X9636PQ support the 4x 10GE break-out mode to operate as 4 individual 10GE ports. This allows the line card to provide up to 144 10GE SFP+ ports.

This line card features a PHY-less design. This reduces the data transport latency on the port by 100ns, decreases the port power consumption and improves reliability due to fewer active components.

The trace lengths from each NFE to the 12 QSFP optics it supports are all under 7”, alleviating the need for re-timers. This further simplifies the line card design and reduces the number of active components.

Figure 8. Nexus 9500 Series 36x 40GE QSFP Line Card



48x 1/10G SFP+ Line Card (N9K-X9564PX)

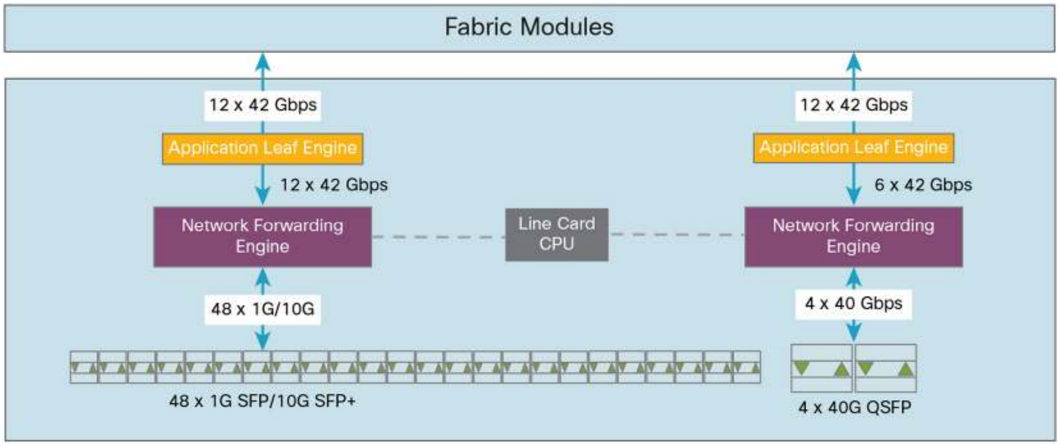
N9K-X9564PX (Figure 9) is an ACI ready leaf line card. It provides 48 1GE SPF/10GE SPF+ ports and 4 40GE QSFP ports. Each of its 4 40GE ports supports the 4x10GE break-out mode to function as 4 individual 10GE ports. As a result, the line card can provide a maximum of 64 10GE ports. The port speed flexibility allows simple and cost-effective network access and aggregation design.

The key components on this line cards include two NFEs, two ALEs, and a line card CPU. The two NFEs provide the front panel ports. One NFE has 48 1/10G ports and the other has 4 40G ports. The two ALEs provide extended buffer space, additional packet handling and the option to use the line card in ACI mode.

To provide port type and speed flexibility, the front panel ports of this line card can operate at different speeds. Port speed mismatch is one of the primary reasons for port congestion and packet buffering. Consequently, this line card may need more buffer space than what its NFEs can provide. The two ALEs provide up to 40MB of additional buffering each. Since ALE is located between NFEs and fabric modules, it can buffer transit traffic between them. Locally switched traffic from a 10G port to a 1G port on the same NFE can also be redirected to the ALE sitting on its north bound interface to take advantage of the extended buffer space.

Like N9K-X9636PQ, this line card also benefits from a PHY-less design for lower power consumption, lower latency as well as higher reliability.

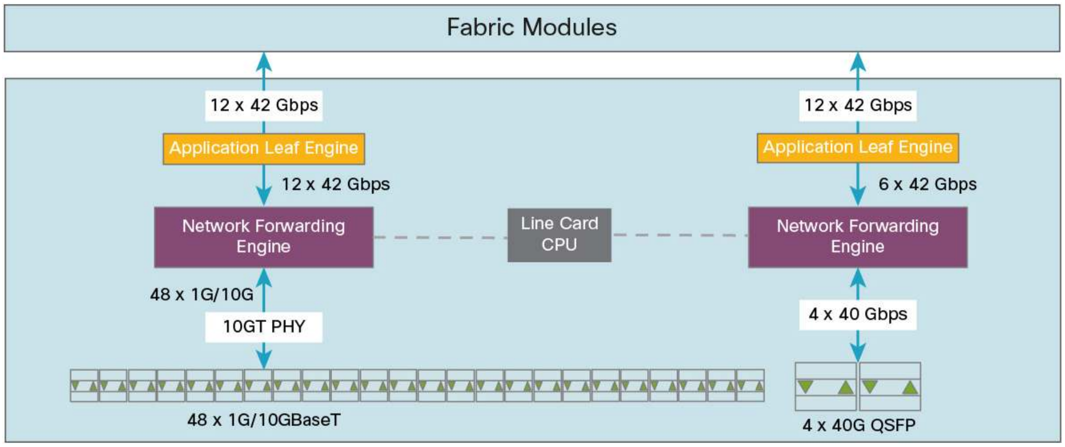
Figure 9. Nexus 9500 Series 48x 1/10GE SPF+ & 4x 40GE QSFP Line Card



48x 1/10G BastT Line Card (N9K-X9564TX)

N9K-X9564TX (Figure 10) is another ACI ready leaf line card. It provides 48 1G/10GBaseT ports and 4 40G QSFP ports. It has an architecture similar to N9K-X9564PX except that all the 48 1G/10GBasteT ports are implemented with 10GT PHYs to convert to 1G/10GBaseT physical media.

Figure 10. Nexus 9500 Series 48x 1/10GBaseT & 4x 40GE QSFP Line Card



Nexus 9500 Series Unicast Packet Forwarding

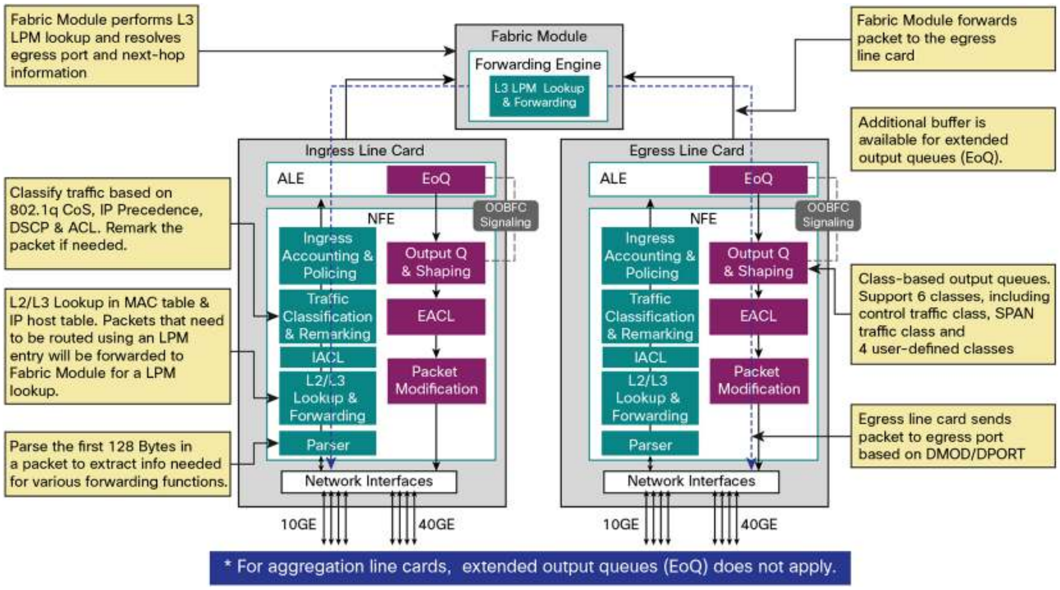
As aforementioned, both line cards and fabric modules on Nexus 9500 Series switches have NFEs that perform packet lookup and forwarding functions. Each NFE has forwarding table resources, which include TCAM tables and a programmable hash table known as Unified Forwarding Table (UFT). It can be flexibly allocated for L2 MAC entries, IP Host entries or LPM entries. This flexibility together with the fully distributed data forwarding architecture allows the Cisco Nexus 9500 Series Switches to optimize the table resource utilization on the line cards and the fabric modules to maximize the Layer-2 and Layer-3 forwarding scalability of the system. It also provides the ability to deploy the Nexus 9500 switches in a broad range of data center scales with a variety of application types.

	Line Card	Fabric Module
L2 MAC Table	160K	-
L3 Host Table	88K	-
LPM Table	-	128K

The data plane forwarding architecture of the Cisco Nexus 9500 Series Switches includes the ingress pipeline on the ingress NFE, fabric module forwarding, and the egress pipeline on the egress NFE. The ingress and egress pipelines may be run on the same line card, or even the same NFE, if the ingress and egress ports are on the same NFE.

A NFE consists of an ingress processing pipeline, a buffer manager for queuing and scheduling, and an egress processing pipeline. The ingress processing pipeline performs packet header parsing, tunnel termination, VRF detection, L2/L3 lookup based on the information in the parsed packet header, and ingress ACL processing. The buffer manager is responsible for all queuing and scheduling functions. The egress pipeline handles all packet modification and egress ACLs. All lookups such as L2/L3/ACL tables are done at the ingress pipeline. Both ingress and egress pipelines have multiple stages to allow for parallel processing of the packets.

Figure 11. Nexus 9500 Unicast Packet Forwarding


Fabric Module performs L3 LPM lookup and resolves egress port and next-hop information
Fabric Module
Forwarding Engine
L3 LPM Lookup & Forwarding
Fabric Module forwards packet to the egress line card
Additional buffer is available for extended output queues (EoQ).
Classify traffic based on 802.1q CoS, IP Precedence, DSCP & ACL. Remark the packet if needed.
L2/L3 Lookup in MAC table & IP host table. Packets that need to be routed using an LPM entry will be forwarded to Fabric Module for a LPM lookup.
Parse the first 128 Bytes in a packet to extract info needed for various forwarding functions.
Ingress Line Card
ALE
EoQ
NFE
OOBFC Signaling
Ingress Accounting & Policing
Output Q & Shaping
Traffic Classification & Remarking
EACL
IACL
L2/L3 Lookup & Forwarding
Packet Modification
Parser
Network Interfaces
10GE
40GE
Egress Line Card
ALE
EoQ
NFE
OOBFC Signaling
Ingress Accounting & Policing
Output Q & Shaping
Traffic Classification & Remarking
EACL
IACL
L2/L3 Lookup & Forwarding
Packet Modification
Parser
Network Interfaces
10GE
40GE
Class-based output queues. Support 6 classes, including control traffic class, SPAN traffic class and 4 user-defined classes
Egress line card sends packet to egress port based on DMOD/DPORT
* For aggregation line cards, extended output queues (EoQ) does not apply.

1. Ingress Processing Pipeline

Packet Header Parsing

When a packet ingresses through a front panel port, it goes through the ingress pipeline on the network forwarding engine of the line card. The first step is packet header parsing. The flexible packet parser parses the first 128 bytes of the packet to extract and save the information such as L2 header, EtherType, L3 header, TCP IP protocols. This is used for subsequent packet lookup and processing logic.

L2 MAC and L3 Host Lookup

As the packet goes through the ingress pipeline, it is subject to L2 switching and L3 routing lookups. First, the NFE examines the destination MAC address (DMAC) of the packet to determine if the packet needs to be L2 switched or L3 routed. If the DMAC matches the switch's own router MAC address, the packet is passed to the L3 routing lookup logic. If the DMAC doesn't belong to the switch, a L2 switching lookup based on the DMAC and VLAN ID is performed. If a match is found in the MAC address table, the packet is sent towards the egress port. If there's no match for DMAC and VLAN combination, the packet is forwarded to all ports in the same VLAN.

As part of the L2 switching logic, the NFE also performs source MAC (SMAC) address lookup for hardware-based learning. The SMAC along with the VLAN ID is used to search the MAC address table. If there is no match, this new address is learned and associated with the ingress port of the packet. If a match is found, no learning action is performed. The NFE supports hardware assisted aging as well. Entries that are not used for an extended period of time (a configurable aging time) are automatically deleted.

Inside the L3 lookup logic on the line card NFE, the destination IP address (DIP) is used for search in the L3 host table. This table stores forwarding entries for directly attached hosts or learned/32 host routes. If the DIP matches an entry in the host table, the entry indicates the destination port, next-hop MAC address and egress VLAN. If there is no match to the DIP in the host table, the packet will be forwarded to fabric module where the longest prefix match (LPM) lookup is performed in the LPM routing table.

When performing Layer-2 switching and Layer-3 host routing, if the egress port is local to the NFE, packets will be forwarded locally by the NFE without going to fabric modules. In the case of ACI ready leaf line card, if the ingress port has a higher speed than the egress port, packets are redirected to the Application Leaf Engine (ALE) for additional buffering in order to compensate for the port speed mismatch.

Ingress ACL Processing

In addition to forwarding lookups, the packet undergoes ingress ACL processing. The ACL TCAM is checked for ingress ACL matches. Each NFE has an ingress ACL TCAM table of 4K entries to support system internal ACLs and user-defined ingress ACLs. These ACLs include Port ACLs, Routed ACLs, and VLAN ACLs. ACL entries are localized to the NFE, and are only programmed where needed. This allows the maximum utilization of ACL TCAM within a Nexus 9500 switch.

Ingress Traffic Classification

Nexus 9500 Series Switches support ingress traffic classification. On an ingress interface traffic can be classified based on the address fields, 802.1q CoS, and IP Precedence or DSCP in the packet header. The classified traffic can be assigned to one of the four qos-groups. The qos-groups serve as an internal identification of the traffic classes that is used for the subsequent QoS processes as packets go through the system.

Ingress Admission, Queuing and Policing

The buffer manager performs ingress accounting and admission functions on the traffic in the ingress processing pipeline. Each NFE has a 12 MB buffer that consists of 60K 208-Byte cells. This buffer resource is dynamically shared by ingress and egress traffic. The ingress admission control mechanism decides if a packet should be admitted into the memory. This decision is based on the amount of buffer memory available and the amount of buffer already utilized by the ingress port and traffic class.

The Nexus 9500 Series Switches supports ingress class-based policing. Policing policies can be defined using a one rate and two colors mechanism or a two rate and three colors mechanism.

2. Fabric Module LPM Lookup

When a packet is forwarded to a fabric module, the fabric module will take different actions based on the lookup results on the ingress line card. In cases where packet is an L2 switched or L3 host routed packet, the ingress line card has resolved the egress port, the next-hop MAC address and the egress VLAN information. The fabric module will simply forward the packet to the egress line card. If the packet needs a LPM lookup, the fabric module searches the LPM table and uses the best match for the destination IP address (DIP) to forward the packet. If there’s no match for the DIP, the packet is dropped. The Unified Forwarding Table (UFT) on the fabric module’s network forwarding engine has a LPM scale of 128K entries.

3. Egress Processing Pipeline

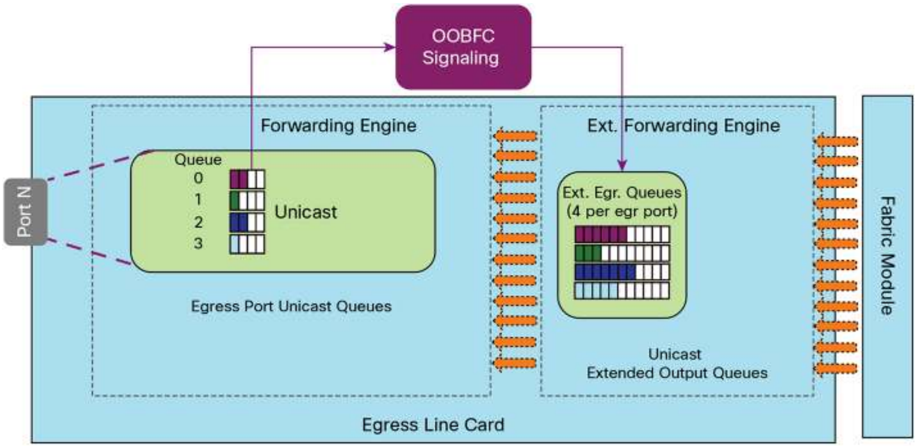
The egress processing pipeline is relatively simple as most lookups and decisions are already made in the ingress pipeline. However, an important function performed in the egress pipeline is egress QoS, including WRED/ECN, egress queuing and shaping.

Egress Queuing and Scheduling

Following the design principle of simplicity and efficiency, Nexus 9500 Series Switches use a simple egress queuing architecture. In the event of egress port congestion, packets are directly queued in the buffer of the egress line card. There are no virtual output queues (VoQ) on the ingress line cards. This greatly simplifies the system buffer management and queuing implementation. A Nexus 9500 switch can support up to six traffic classes on egress (four user defined classes identified by qos-group IDs, a CPU control traffic class and a SPAN traffic class). Each user defined class can have a unicast queue and a multicast queue per egress port. The 12 MB buffer on a NFE is shared among the local ports. The switch software has a mechanism to meter and limit buffer utilization per egress port. This ensures that no single port can consume more than its fair share of the buffer memory leading to buffer starvation of other ports.

The ACI ready leaf line cards have an additional 40MB buffer in each of their ACI Leaf Engines(ALE). 10MB of the buffer is allocated to fabric bound traffic. The remaining 30MB is allocated to egress traffic from fabric modules and locally switched traffic from higher speed ingress port to a lesser speed egress port. This 30MB buffer is used for extended output queues for unicast traffic. The NFE communicates the unicast queue status to the ALE through an out-of-band flow control (OOBFC) signaling channel. When an egress queue exceeds the configured threshold, the NFE sends an OOBFC signal to instruct the ALE to stop forwarding traffic for this queue and start queuing packets in its own buffer. Upon receiving this signal, the ALE starts to form the extended output queue for this traffic class on the given egress port. When the egress queue length is reduced to the configured restart threshold, the NFE sends another OOBFC signal to direct the ALE to resume transmitting traffic for this particular queue.

Figure 12. Nexus 9500 Extended Output Queue (EoQ)



The egress queuing architecture with extended output queues, though simple, is a highly efficient approach to deal with port congestion with fairness. It ensures that no one port can cause buffer starvation of any other port.

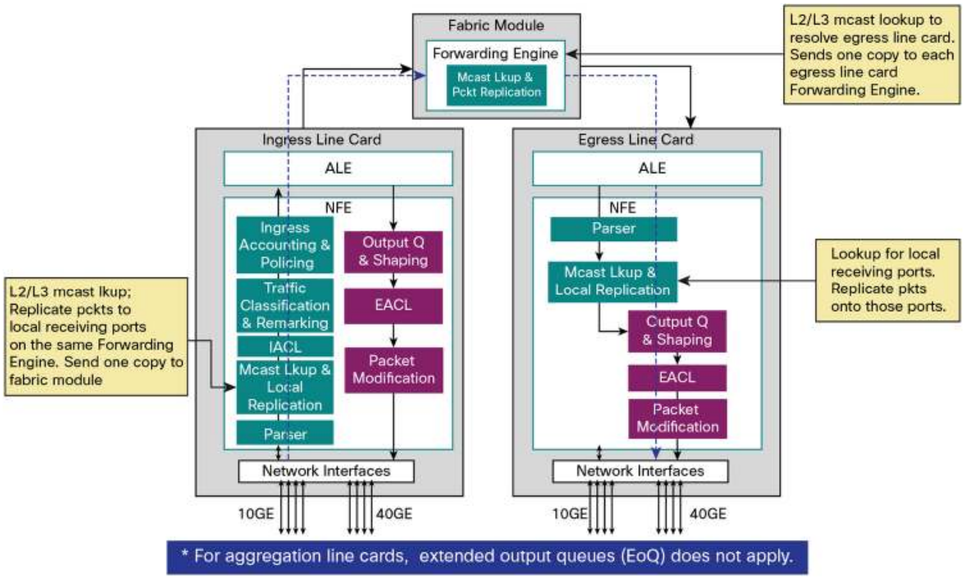
Nexus 9500 Series Multicast Packet Forwarding

Multicast packets go through the same ingress and egress processing pipelines as the unicast packets. However, one difference in packet lookup & forwarding process is, the Nexus 9500 switches perform 3-stage distributed multicast lookup and replication. The multicast routing table is stored on all line cards and fabric modules. The ingress NFE performs the 1st lookup to resolve local receivers. If there are any local receivers, the NFE creates one copy per local receiving port. Also, the ingress NFE sends a copy of the incoming packet to the fabric module.

On receiving the packet, the fabric module does the 2nd lookup to find the egress line cards. The fabric module replicates the packet to each egress NFE.

The egress NFE does the 3rd lookup to resolve its local receivers and replicates the packet onto those ports. This multi-stage multicast lookup and replication is the most efficient way of replicating and forwarding multicast traffic.

Figure 13. Nexus 9500 Multicast Packet Forwarding


Fabric Module
Forwarding Engine
Mcast Lkup & Pckt Replication
L2/L3 mcast lookup to resolve egress line card. Sends one copy to each egress line card Forwarding Engine.
Ingress Line Card
ALE
NFE
Ingress Accounting & Policing
Traffic Classification & Remarking
IACL
Mcast Lkup & Local Replication
Parser
Output Q & Shaping
EACL
Packet Modification
Network Interfaces
10GE
40GE
L2/L3 mcast lkup; Replicate pckts to local receiving ports on the same Forwarding Engine. Send one copy to fabric module
Egress Line Card
ALE
NFE
Parser
Mcast Lkup & Local Replication
Output Q & Shaping
EACL
Packet Modification
Network Interfaces
10GE
40GE
Lookup for local receiving ports. Replicate pkts onto those ports.
* For aggregation line cards, extended output queues (EoQ) does not apply.

Another difference between multicast and unicast traffic forwarding is that there are no extended output queues for multicast traffic. Network forwarding engine supports four multicast queues per egress port. In the presence of ACI Leaf Engines, it queues multicast traffic independently to the network forwarding engine multicast queues. There is no backpressure signal to control the multicast queues through the OOBFC channel.

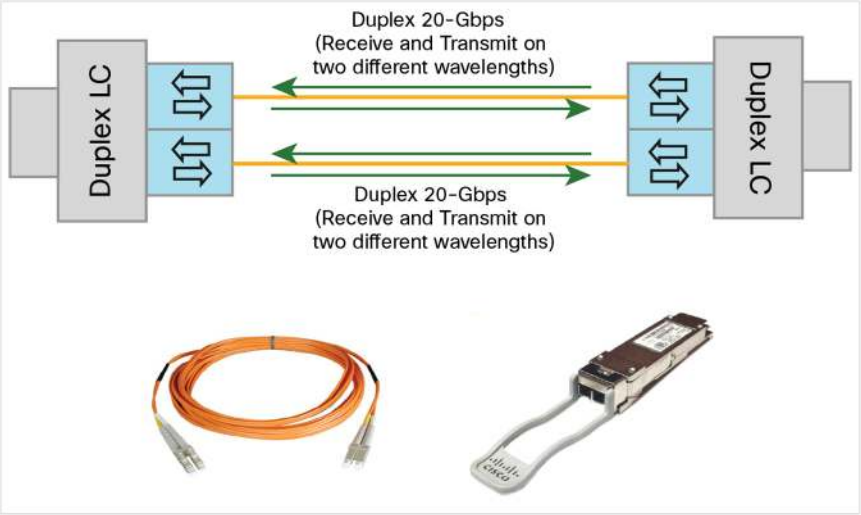
Cisco QSFP Bi-Di Technology for 40Gbps Migration

With its high port density and performance for 1/10/40 GE connectivity, Nexus 9500 Series Switches cater to the next generation data center infrastructure. While offering 1/10GE at the access/leaf and 40GE links at the aggregation/spine, it provides more scalable bandwidth for datacenter applications.

However, migrating an existing data center network from 10GE to 40GE involves more than a network platform upgrade. Cabling infrastructure migration is one of the biggest challenges in this task. The current 10GE cabling infrastructure uses 2 MMF fiber strands for one 10GE connection. However, the existing short reach 40GE optics transceivers, either SR4 or CSR4, feature independent transmitter and receiver sections, each with 4 fiber strands in parallel. As a result, 8 fiber strands are required for a duplex 40GE connection. These differences dictate that moving the current 10GE infrastructure to 40GE using the existing 40GE optical transceivers requires a forklift cabling infrastructure upgrade or rebuild. The staggering cost and the potential service interruption makes it very difficult to migrate an existing production data center to a 40GE infrastructure.

Cisco QSFP Bi-Directional transceiver technology solves this problem by providing the capability of transmitting full duplex 40G over two strands of MMF fiber with LC connectors. In other words, the QSFP BiDi transceiver allows 40GE connectivity to re-use the existing 10G fibers and fiber trunk without the need of expansion or rebuild. It also removes 40-Gbps cabling cost barriers for migration from 10-Gbps to 40-Gbps connectivity in data center networks.

Figure 14. Cisco BiDi Transceiver Technology



Conclusion

Nexus 9500 Series Switches are the industry leading data center class switches offering the highest port density for 1/10/40, and in the future, 100GE connectivity, in addition to the unprecedented true line rate and low latency forwarding performance. Nexus 9500 Series Switches support the industry leading 10GE and 40GE port density. With its flexible port speeds and chassis form factors, Nexus 9500 Series Switches are catering to virtualized, multitenant and cloud data center deployments at scales from small to medium to large.

The midplane-free chassis design enables the maximum cooling efficiency. The combination of merchant and custom silicon allows the line cards to have the fewest number of ASICs while providing a record-setting performance. With innovations such as, the front-to-rear airflow, and the 80PLUS Platinum certified highly efficient power supplies, Nexus 9500 Series Switches set a new record for energy efficiency, reliability and performance for data center class switches.

By decoupling the intra-system management from the switch control plane, the Nexus 9500 Series Switches achieve a control plane of unprecedented stability. Equipped with a supervisor engine built with the latest multicore CPU, together with line card CPUs to offload tasks from the supervisor engines, Nexus 9500 Series Switches provide the foundation of a reliable data center switch.

Operating in the classic NX-OS mode, Nexus 9500 Series Switches run on a single image for all switches in the family, greatly simplifying network administration. While running on the latest 64-bit Linux Kernel with true process modularity, high software resiliency together with multiple enhancements in automation and programmability, the enhanced NX-OS for Nexus 9500 Series Switches is the best solution for data centers that seeks to modernize and automate data center network management and operating models.

With the above mentioned unique features, the Cisco Nexus 9500 Series Switches are the ideal data center switches to allow organizations to build a reliable, scalable, resilient and automated data centers.

Appendix

Appendix A - Terminologies

ACI - Application Centric Infrastructure

NFE - Network Forwarding Engine

ALE - ACI Leaf Engine

EoQ - Extended Output Queue

OOBFC - Out-of-Band Flow Control

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-729987-00 11/13